Попов Сергей Витальевич
кандидат технических наук,
зав. сектором наукометрии и статистики науки РИЭПП.
Тел. (495) 916-14-79,
info@riep.ru

ТЕМАТИЧЕСКИЙ ПОИСК В ИНТЕРНЕТЕ: НАЗАД В БУДУЩЕЕ

В основе современных поисковых машин, работающих в Интернете, лежат алгоритмы документального поиска, разработанные ещё в 60-х — 70-х годах прошлого века. В то время документальные информационно-поисковые системы (ИПС) в первую очередь предназначались для поиска научно-технической информации и основными их пользователями были инженеры и ученые. Вот что пишет в связи с этим разработчик системы Яндекс Илья Сегалович: «Но что же поменялось в действительности за последние годы? Не алгоритмы и не структуры данных, не математические модели. Хотя и они тоже. Поменялась парадигма использования систем. Проще говоря, к экрану со строчкой поиска подсели домохозяйка, ищущая утюг подешевле, и выпускник вспомогательного интерната в надежде найти работу автомеханика» [1].

Другими словами, подключение к поиску информации широких масс населения коренным образом повлияло на развитие ИПС. Да, основные поисковые алгоритмы остаются прежними. Но, согласитесь, что поиск расписания электропоездов и поиск информации о рентгенолитографии — две разные задачи. К сожалению, задачи тематического научного поиска уходят на периферию интересов разработчиков популярных поисковиков Интернета. Так, Илья Сегалович пишет: «Мечты 60-х — 80-х об итеративном уточнении запросов, о понимании естественного языка, о поиске по смыслу, о генерации связного ответа на вопрос с трудом выдерживают сейчас жестокое испытание реальностью» [1].

С другой стороны, небескорыстное желание владельцев известных поисковых машин привлечь как можно больше разнообразных пользователей приводит к новым проблемам. Как сказано в работе [2], все основные технологические проблемы Интернета, которые мы сейчас видим и зачастую ощущаем на себе, имеют своей причиной то, что когда эти технологии разрабатывались, никто из разработчиков (по их собственным признаниям) не представлял себе, что Интернет станет глобальной информационной средой.

Особое внимание следует уделить алгоритмам ранжирования выдачи (ранжирование по релевантности).

Словарные ИПС способны выдавать списки документов, содержащие миллионы ссылок. Даже просто просмотреть такие списки невозможно, да и не нужно. Было бы удобно иметь возможность задать формальные критерии (хотя бы относительной) важности документов, с тем, чтобы наиболее важные документы попадали бы в начало списка. Все разра-

ботчики ИПС в настоящее время уделяют основное внимание именно алгоритму ранжирования полученных ссылок.

Наиболее часто используемыми критериями при ранжировании в поисковых машинах Интернета являются:

– наличие слов из запроса в документе, их количество, близость к началу документа, близость друг к другу;

– наличие слов из запроса в заголовках и подзаголовках документов (заголовки должны быть специально отформатированы);

– количество ссылок на данный документ с других документов;

– «респектабельность» ссылающихся документов.

Как видно из критериев ранжирования, реальный критерий релевантности документа — наличие слов из запроса — не так сильно влияет на его ранг в результатах поиска. С другой стороны, использование синтетических критериев дает возможность манипулирования результатами вычислений ранга страницы, с чем и борются все ИПС. Такая ситуация ведет к снижению качества поиска, поскольку потенциально более полезные документы неминуемо оттесняются своими «оптимизированными» конкурентами в конец списка. Наверно, многие сталкивались с тем, что реально полезные ресурсы в поисковиках находятся на второй-третьей странице выдачи поискового запроса [2]. В случае тематического поиска релевантные ссылки могут находиться и на 10-й, и даже на 100-ой страницах выдачи.

Интересный факт сообщает новостная служба портала «Открытые системы».

Как показало исследование, проведенное специалистами Квинслендского технологического университета и Университета Пенсильвании с помощью портала метапоиска Dogpile.com, крупнейшие поисковые системы крайне редко выдают идентичный набор верхних строчек результатов при поиске по одним и тем же запросам. Исследователи в общей сложности ввели около 19,3 тыс. запросов к Google, Yahoo, Windows Live Search и Ask.com. Совпадение первого результата во всех системах было выявлено только в 3,6% случаев. Совпадение первых трех не выпало ни разу, даже если не учитывать порядок следования результатов. В среднем менее 1% результатов первой страницы совпадало на всех четырех сайтах. Для сравнения, четыре года тому назад в аналогичном исследовании совпадения первого результата отмечались в 7% случаев [3].

В то же время, по данным компании Forrester Research [4]:

- 90% пользователей находят новые сайты через поисковые системы;
- работа с поисковыми системами — второй по популярности вид деятельности в Интернете после использования электронной почты;
- 80% пользователей поисковых систем не смотрят результаты дальше первой страницы;
- по сравнению с баннерной рекламой, посетители сайта в пять раз охотнее станут вашими клиентами, найдя ваш сайт через поисковую систему;
- 55% онлайн покупок и заказов совершаются на сайтах, найденных через поисковые системы;

- четверо из пяти пользователей используют поисковые системы ежедневно.

Другими словами, поисковые машины — один из самых важных инструментов работы с информацией в Интернете. Мы уже частично дали ответ на вопрос, почему поисковые машины Интернета не используют весь арсенал средств поиска документальной информации, разработанных во второй половине прошлого века. Одна из основных причин — смена категорий пользователей, а следовательно, смена типов запросов. С другой стороны, в настоящее время, судя по спискам использованных источников в статьях и книгах, ученые все чаще обращаются в глобальную сеть для поиска необходимой им информации. Таким образом, тематические запросы снова становятся предметом головной боли разработчиков поисковых машин, и, следовательно, они вынуждены обращаться к опыту прошлых лет. Как пишет Сегалович, «все многообразие моделей традиционного информационного поиска принято делить на три вида: теоретико-множественные (булевская, нечетких множеств, расширенная булевская), алгебраические (векторная, обобщенная векторная, латентно-семантическая, нейросетевая) и вероятностные. Булевское семейство моделей, по сути, — первое, приходящее на ум программисту, реализующему полнотекстовый поиск. Есть слово — документ считается найденным, нет — не найденным» [1].

Далее приведены некоторые этапы развития моделей документального поиска:

1. 1957 год. Т. Джойс и Р.М. Нидхэм предложили векторную модель поиска.
2. 1960 год. М.Е. Марон и Дж.Л. Кунс предложили вероятностную модель поиска.
3. 1968 год. Векторная модель реализована Герардом Сэлтоном (Gerard Salton) в поисковой системе SMART (Salton’s Magical Automatic Retriever of Text).
4. 1977 год. К.Е. Робертсон и К. Спарк-Джоунз обосновали и реализовали вероятностную модель поиска.
5. 1988 год. Дж.В. Фурнас и С.Дирвестер разработали метод латентно-семантического индексирования.

Опыт моделирования документального поиска, накопленный в прошлом веке, постепенно начинает использоваться при разработке поисковых машин Интернета. Среди отечественных Интернет-поисковиков, использующих такой опыт, можно отметить системы Галактика-Зум (корпорация Галактика), Артефакт (компания Интегрум-Техно), Nigma (МГУ).

Как уже отмечалось выше, одной из основных проблем поисковых систем в Интернете является неэффективность алгоритмов ранжирования найденных документов. Это во многом обусловлено тем, что поисковые запросы в среднем состоят всего из двух-трёх слов, т. е. просто не хватает исходной информации для эффективного ранжирования выдачи. В уже упомянутой ИПС SMART проблема, связанная с короткими запросами, была успешно преодолена с помощью так называемой «об-

ратной связи по релевантности». При этом поиск проходит в несколько итераций. На каждом шаге итерации поисковый запрос расширяется за счет терминов, выделенных пользователем из понравившихся ему среди найденных на этом шаге документов. Заметим, что сам термин «ранжирование по релевантности» появился на фоне реализации обратной связи по релевантности в системе SMART [5].

Попытки реализации обратной связи по релевантности в Интернете осуществляются, например, в отечественной поисковой системе WEB ИРБИС, работающей с массивами научной информации (ИНИОН, ГПНТБ).

В заключение хочется отметить, что противоречие между коммерциализацией и качеством поиска ИПС в глобальной компьютерной сети продолжает существовать.

Вот, например, еще одна новость с портала «Открытые системы».

В компании Yahoo надеются, что с переходом на новую поисковую технологию ей удастся восстановить позиции на рынке, где сейчас преобладает Google. В числе улучшений — упрощенный пользовательский интерфейс с меньшим количеством баннеров, функция поиска изображений и модификация настроек (выбор одного из 30 поддерживаемых языков), поиск с учетом домена, страны и времени создания документов. Служба расположена по новому адресу: http://new.search.yahoo.com. Поисковая система основана на усовершенствованном варианте технологии компании Inktomi, приобретенной Yahoo. Кроме того, система частично полагается на технологии Google. *Привлечь внимание к порталу в Yahoo рассчитывают за счет совершенствования его служб, в числе которых — спортивные результаты, желтые страницы, поиск по Internet-магазинам, знакомства, биржа труда и т. д. [6].*

Литература

1. *Сегалович И. В.* Как работают поисковые системы // Мир Internet. 2002. № 10. С. 24—32.
2. *Тактаев Станислав.* Поиск информации в компьютерных сетях: новые подходы. http://www.searchengines.ru/articles/004603.html.
3. http://www.osp.ru/news/2007/0618/4233621/.
4. http://promo.by/.
5. *Сэлтон Г.* Автоматизированная обработка, хранение и поиск информации. М.: Советское радио, 1973.
6. http://www.osp.ru/news/2003/0410/611142.